Statistical Inference

Frank Schorfheide[1]

This Version: October 8, 2018

Abstract

Joint, conditional, and marginal distributions for data and parameters, prior distribution, posterior distribution, likelihood function, loss functions, decision-theoretic approach to inference, frequentist risk, minimax estimators, admissible estimators, posterior expected loss, integrated risk, Bayes risk, calculating Bayes estimators, generalized Bayes estimators, improper prior distributions, maximum likelihood estimators, least squares estimators. Neyman-Pearson tests, null hypothesis, alternative hypothesis, point hypothesis, composite hypothesis, type-I error, type-II error, power, uniformly most powerful tests, Neyman-Pearson lemma, likelihood ratio test, frequentist confidence intervals via inversion of test statistics, Bayesian credible sets.

1 Introduction

After a potted introduction to probability we proceed with statistical inference. Let $\mathcal{X}$ be the sample space and Θ be the parameter space. We will start from the product space $\mathcal{X} \otimes \Theta$. We equip this product space with a sigma-algebra as well as a joint probability distribution $\mathbb{P}^{X,\theta}$. We denote the marginal distributions of X and θ by

$$\mathbb{P}^X, \quad \mathbb{P}^\theta$$

and the conditional distributions by

$$\mathbb{P}^X_\theta, \quad \mathbb{P}^\theta_X,$$

where the subscript denotes the conditioning information and the superscript the random element. In the context of Bayesian inference the marginal distribution $\mathbb{P}^\theta$ is typically called a *prior* and reflects beliefs about θ before observing the data. The conditional distribution $\mathbb{P}^\theta_X$ is called a *posterior* and summarizes beliefs after observing the data, i.e., the realization of the random variable X. The most important object for frequentist inference is $\mathbb{P}^X_\theta$, which is the sampling distribution of X conditional on the unknown parameter θ.

[1]Department of Economics, PCPSE Room 621, 133 S. 36th St, Philadelphia, PA 19104-6297, Email: *schorf@ssc.upenn.edu*, URL: *https://web.sas.upenn.edu/schorf/*.

If we have densities (w.r.t. Lebesgue measure) we use the notation

$$p(x), \quad p(\theta), \quad p(x|\theta), \quad p(\theta|x).$$

As a function of θ, the function $\mathcal{L}(\theta|x) = p(x|\theta)$ is called *likelihood function.* It indicates how "likely" the realization $X = x$ is as a function of the parameter θ. The basic problem of statistical inference is that θ is unknown. The econometrician/statistician observes a particular realization of $X = x$ and wants to make statements about θ.

Example: Throughout this lecture, we consider a simple Gaussian location problem. Let

$$\theta \sim N(0, 1/\lambda), \quad X|\theta \sim N(\theta, 1). \tag{1}$$

The econometrician observes a realization x of the random variable X. Note that we only have a single observation in our sample. □

2 Point Estimation

2.1 Methods For Generating Estimators

Point estimators are mappings from the sample space $\mathcal{X}$ into the parameter space Θ. There are many ways of constructing these mappings. In the context of (1), many estimators will look identical, but in the context of richer models, they will look different, and have different properties. Throughout this section we will use upper case letters for random variables, i.e., X and lower case letter, i.e., x, for realizations of random variables. An estimator can be both. Whenever we study its probability properties, we treat it as a random variable. Whenever we compute based on an observed sample, we regard it as a realization of a random variable.

Least Squares. The ordinary least squares (OLS) estimator can be defined as follows. Rewrite the model as

$$X = \theta + U, \quad U|\theta \sim N(0, 1).$$

Now define the estimator as

$$\hat{\theta}_{LS} = \text{argmin}_{\theta \in \Theta} \, (x - \theta)^2. \tag{2}$$

In this example $\hat{\theta}_{LS} = x$.

Maximum Likelihood. The maximum likelihood estimator (MLE) is obtained by maximizing the likelihood function. Rather than maximizing the likelihood function, it is typically

more convenient to maximize the log-likelihood function. The log is a monotone transformation which preserves the extremum. Let

$$\ell(\theta|x) = \ln \mathcal{L}(\theta|x) = \ln p(x|\theta) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}(x-\theta)^2.$$

Thus, the MLE is

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta\in\Theta}\, \ell(\theta|x), \tag{3}$$

which leads to $\hat{\theta}_{MLE} = x$.

Method of Moments Estimator. Methods of moments estimators match model-implied moments, which are functions of the parameter with sample moments. In our model

$$\mathbb{E}_\theta^X[X] = \theta.$$

In the data, we can approximate the expected value with the sample mean. Because we only have a single observation, the sample mean is trivial $\bar{x} = x$:

$$\widehat{\mathbb{E}_\theta^X}[X] = \bar{x} = x.$$

Now we minimize the discrepancy between the model-implied moment and the sample moment:

$$\begin{aligned}\hat{\theta}_{MM} &= \text{argmin}_{\theta\in\theta}\left(\mathbb{E}_\theta^X[X] - \widehat{\mathbb{E}_\theta^X}[X]\right)^2 \\ &\quad \text{argmin}_{\theta\in\theta}\,(\theta - x)^2.\end{aligned} \tag{4}$$

Thus, we obtain $\hat{\theta}_{MM} = x$.

Bayes Estimator. The derivation of Bayes estimators requires the calculation of posterior distributions. According to Bayes Theorem, the conditional density of θ given the observation x is given by

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta).$$

Here $p(\theta)$ is called the prior density and captures information about θ *prior* to observing the data x. If λ is small, then the prior density has a large variance and is very "diffuse." If λ is large, then the prior density has small variance and the prior is very concentrated. A small λ prior is often called uninformative, and a large λ prior is very informative.

In the Gaussian shift experiment

$$p(x|\theta)p(\theta) \propto \exp\left\{-\frac{1}{2}(x-\theta)^2\right\}\left\{-\frac{1}{2}\lambda\theta^2\right\}$$

Re-arranging terms, we find that

$$p(x|\theta)p(\theta) \propto \exp\left\{-\frac{\lambda+1}{2}\left(\theta - \frac{1}{\lambda+1}x\right)^2\right\}$$

which implies that

$$\theta|x \sim N\left(\frac{1}{\lambda+1}x, \frac{1}{\lambda+1}\right).$$

In a last step we need to turn the posterior density into a point estimator. For now, let's just consider the mean of the posterior distribution (we will discuss conditions under which this is a reasonable choice below):

$$\hat{\theta}_B = \mathbb{E}_X^{\theta}[\theta] = \int \theta p(\theta|x)d\theta = \frac{1}{\lambda+1}x. \tag{5}$$

Note that the posterior mean is in absolute value smaller than x. Here the MLE is pulled toward the prior mean, which is zero. In this sense, the prior induces "shrinkage."

A Class of Linear Estimators. Note that all the estimators that we have derived have the form

$$\hat{\theta} = cx.$$

For the OLS, ML, and MM estimators $c = 1$. For the Bayes estimator, depending on the choice of λ, $0 \leq c \leq 1$.

2.2 Evaluation of Estimators

As we have seen, there are many ways of constructing point estimators. We now need a way of distinguishing good from bad estimators. To do so, we will adopt a decision-theoretic approach. Let $\delta \in \mathcal{D}$ be a "decision" and $\mathcal{D}$ the decision space. In our context, the decision is to report a point estimate of θ. Thus, $\mathcal{D} = \mathbb{R}$. Given the setup of our experiment the probability that $\delta = \theta$ is zero. Thus, we need a loss function that weighs the discrepancy between δ and θ. Mathematically one of the most convenient loss functions is the quadratic loss function:

$$L(\theta, \delta) = (\theta - \delta)^2.$$

The decision, i.e., the estimator of θ, can and should depend on the random variable X. We write $\delta(X)$ to denote this dependence and to highlight that the decision, from a pre-experimental perspective, is a random variable. In general, our goal will be to choose a function $\delta(X)$ such that the *expected loss* is small. In slight abuse of notation we will denote the domain of the functions $\delta(X)$ also by $\mathcal{D}$.

2.3 Frequentist Risk

The *expected loss* is called *risk*. Since we are working on a product space, we can take expectations with respect to different measures. We will start by taking expectations with respect to the sampling distribution $\mathbb{P}_\theta^X$. This type of analysis is often called *frequentist* or *classical*. Basically, it tries to determine the behavior of an estimator conditional on a "true" θ under the assumption that *nature* provides the econometrician repeatedly with draws from the random variable X. This is not a very compelling assumption, but it is quite popular. The frequentist risk is given by

$$\begin{aligned} R(\theta, \delta(\cdot)) &= \mathbb{E}_\theta^X[L(\theta, \delta(\cdot))] \\ &= \int_{\mathcal{X}} L(\theta, \delta(\cdot)) p(x|\theta) dx \end{aligned}$$

Example: In the Gaussian shift experiment we could consider estimators of the form $\delta(x) = cx$, where $c \in \mathbb{R}$ is a constant indexing different estimators. Using the quadratic loss function $L(\theta, \delta) = (\theta - \delta)^2$, we obtain (notice we are taking expectations over X:

$$\begin{aligned} R(\theta, \delta(\cdot)) &= \int_{\mathcal{X}} (\theta - cx)^2 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\theta)^2\right\} dx \\ &= \theta^2 - 2c\theta \mathbb{E}_\theta^X[X] + c^2 \mathbb{E}_\theta^X[X^2] \\ &= \theta^2 - 2c\theta^2 + c^2(1+\theta^2) \\ &= \theta^2(1-c)^2 + c^2. \end{aligned}$$

A little numerical illustration is useful.

	$\theta = 0$	$\theta = 1/2$	$\theta = 1$
$c = 0$	0	1/4	1
$c = 1/2$	1/4	5/16	1/2
$c = 1$	1	1	1

Notice that there is no unique ranking of estimators (independent of the "true" θ), i.e., choices of c. Thus, we need to define some concepts that allow us to rank estimators. □

Definition 1 *The minimax risk associated with the loss function $L(\theta, \delta)$ is the value*

$$\bar{R} = \inf_{\delta \in \mathcal{D}} \sup_\theta R(\theta, \delta)$$

and a minimax estimator is any estimator δ_0 such that

$$\sup_\theta R(\theta, \delta_0) = \bar{R}.$$

We have essentially characterized the Nash equilibrium in a zero-sum game between *nature* and the *econometrician*. Nature gets to choose θ and the econometrician gets to choose δ. The minimax estimator could in principle be a randomized estimator. For instance, in the context of our shift experiment let $V \sim U[0,1]$ and define

$$\delta(X,V) = cX\mathbb{I}\{V \leq 1/2\}.$$

Thus, with 50% probability one sets the estimator to cX and with 50% probability one sets it to zero.

Example: Consider our restricted class of estimators $\mathcal{D} = \{\delta \mid \delta = cx,\ c \in \mathbb{R}\}$. Recall that

$$R(\theta,\delta) = \theta^2(1-c)^2 + c^2.$$

For $c \neq 1$ *nature* can choose a large value of θ, which leads to a large risk $R(\theta,\delta) >> 1$. Thus, the optimal choice for the econometrician is $c = 1$, which guarantees that $R(\theta,\delta) = \bar{R} = 1$.

For the discrete case (see above) in which $c \in \{0, 1/2, 1\}$ and $\theta \in \{0, 1/2, 1\}$ notice that $(c = 1/2, \theta = 1)$ is a Nash equilibrium of a zero-sum game in which the econometrican's loss equals nature's gain. If the econometrician plays $c = 1/2$ then nature maximizes her payoff by playing $\theta = 1$. Likewise, if nature plays $\theta = 1$ then it is optimal for the econometrician to play $c = 1/2$. □

Definition 2 *An estimator δ_0 is inadmissible if there exists an estimator δ_1 which dominates δ_0, that is, for every θ*

$$R(\theta,\delta_0) \geq R(\theta,\delta_1)$$

and for at least one value θ_0 of the parameter

$$R(\theta,\delta_0) > R(\theta,\delta_1).$$

Otherwise, δ_0 is said to be admissible.

Example: Consider the estimator $\delta_0(x) = 10x$. Then,

$$R(\theta,\delta_0) = \theta^2(1-c)^2 + c^2 \geq c^2 = 100.$$

Alternatively, consider the estimator $\delta_1(x) = x$ with

$$R(\theta,\delta_1) = 1 < 100.$$

Thus, δ_0 is inadmissable.

2.4 Integrated Risk and Bayes Risk

Instead of averaging the loss with respect to the conditional distribution $\mathbb{P}_\theta^X$, we can also average with respect to the distribution $\mathbb{P}_X^\theta$. The *posterior expected loss* is defined as

$$\rho(\mathbb{P}^\theta, \delta|X) = \mathbb{E}_X^\theta[L(\theta, \delta)] = \int_\Theta L(\theta, \delta)p(\theta|x)d\theta.$$

The expected loss depends on the prior distribution $\mathbb{P}^\theta$. Instead of being dependent on the parameter value θ, the posterior expected loss depends on the realization of X. The dependence on X is more appealing than the dependence on θ because the realization of X is observed whereas the parameter θ is not.

Given the joint distribution of X and θ it is also possible to define the *integrated* risk

$$\begin{aligned} r(\mathbb{P}^\theta, \delta(\cdot)) &= \mathbb{E}^{X,\theta}[L(\theta, \delta(X))] \\ &= \int_{\mathcal{X}} \int_\Theta L(\theta, \delta(x))p(x, \theta)d\theta dx \\ &= \int_{\mathcal{X}} \rho(\mathbb{P}^\theta, \delta(x)|x)p(x)dx \\ &= \int_\theta R(\theta, \delta(\cdot))p(\theta)d\theta \end{aligned}$$

Thus, the integrated risk averages the posterior expected loss over X and the frequentist risk over θ. The integrated risk associates a real number, rather than a function of θ, with every estimator and thereby induces a total ordering.

Example: In the location shift experiment we considered the class of estimators $\delta(x) = cx$ and obtained

$$R(\theta, \delta) = \theta^2(1 - c)^2 + c^2.$$

Under the prior distribution $\theta \sim N(0, 1/\lambda)$ the integrated risk becomes

$$r(\mathbb{P}^\theta, \delta(\cdot)) = \frac{1}{\lambda}(1 - c)^2 + c^2.$$

minimizing with respect to c yields that the estimator

$$\delta^* = \frac{1}{1 + \lambda}x \tag{6}$$

minimizes the integrated risk. Thus, our Bayes estimator minimizes the integrated risk! □

More generally, an estimator minimizing the integrated risk can be obtained by selecting, for every $x \in \mathcal{X}$ the value $\delta(x)$ which minimizes the posterior expected loss $\rho(\mathbb{P}^\theta, \delta(x)|x)$.

Definition 3 *A Bayes estimator associated with a prior distribution $\mathbb{P}^\theta$ and a loss function $L(\cdot)$ is any estimator $\delta_{\mathbb{P}^\theta}$ which minimizes $r(\mathbb{P}^\theta, \delta(\cdot))$. For every $x \in \mathcal{X}$ it is given by*

$$\delta_{\mathbb{P}^\theta}(x) = \min_{d \in \mathcal{D}} \rho(\mathbb{P}^\theta, d|x).$$

The value $r(\mathbb{P}^\theta) = r(\mathbb{P}^\theta, \delta_{\mathbb{P}^\theta})$.

It is fairly straightforward to verify that under a quadratic loss function the Bayes estimator is given by the posterior mean

$$\delta_{\mathbb{P}^\theta}(X) = \mathbb{E}^\theta_X[\theta] = \frac{1}{\lambda + 1} X,$$

which is identical to (6). Thus, the Bayes estimator belongs to the family $\mathcal{D} = \{\delta \mid \delta = cx,\ c \in \mathbb{R}\}$, with $c \leq 1$.

Notice that as $\lambda \longrightarrow 0$ the Bayes estimator converges to X. Since we have not defined a convergence concept for probability distributions yet we cannot discuss limits of prior distributions in a rigorous manner. However, as $\lambda \longrightarrow 0$, the prior density becomes more and more spread out and essentially starts to look "uniform" on the real line. The uniform measure for the real line is the Lebesgue measure and has the property that its total mass is infinite. Thus, it is not a *proper* probability measure (which we defined to have total mass one). Notice, however, that we could have carried out the preceding calculations under the improper prior $p(\theta) \propto 1$. Then

$$p(\theta|x) \propto p(x|\theta)p(\theta) \propto p(x|\theta)$$

Under this improper prior, the posterior is exactly proportional to the likelihood function. Interestingly, the resulting posterior is proper because the likelihood function is integrable:

$$\theta|x \sim N(x, 1).$$

Once $p(\theta) \propto 1$, the integrated risk is unbounded (except for some special estimators). However, we might still be able to define a *generalized Bayes estimator* that minimizes the posterior expected risk for every possible realization $x \in \mathcal{X}$. In our example, for each x the estimator $\delta(x) = x$ minimizes the posterior expected risk and is therefore a *generalized* Bayes estimator. Under this improper prior the (generalized) Bayes estimator equals the maximum likelihood estimator and the OLS estimator.

There exists an extensive literature exploring the connections between Bayes estimators, minimax estimators, and admissible estimators. Generally speaking, minimax estimators are often obtained by constructing Bayes estimators from a so-called *least favorable* prior distributions. Moreover, under some regularity conditions Bayes estimators are admissible.

To give you a flavor of these results, consider the following example. Suppose $\Theta = \{1, 2\}$. The frequentist risk for a decision δ can be characterized by two numbers:

$$R(\theta = 1, \delta) = r_1, \quad R(\theta = 1, \delta) = r_2.$$

Define the risk set as

$$\mathcal{R} = \{(r_1, r_2) \,|\, r_1 = R(\theta = 1, \delta),\, r_2 = R(\theta = 2, \delta),\, \delta \in \mathcal{D}\}.$$

Suppose that the risk set $\mathcal{R}$ is compact and convex. The lower boundary, $\Gamma(\mathcal{R})$ comprises the set of admissible decision rules. For every $r \in \Gamma(\mathcal{R})$ there exists a tangent line of the form:

$$p_1 r_1 + p_2 r_2 = k.$$

This implies the decision δ associated with r is a Bayes decision under the prior distribution for which $\mathbb{P}(\theta = 1) = p_1/(p_1 + p_2)$. By construction, for every $r' \in \mathcal{R}$

$$p_1 r_1' + p_2 r_2' \geq k.$$

Thus, we can conclude that the set of Bayes estimators constitutes a *complete class* $\mathcal{C}$: For every $\delta' \in \mathcal{C}^c$ there exists a $\delta \in \mathcal{C}$ that dominates δ'.

3 Testing

Sometimes, though less often than the proliferation of econometric tests might suggest, we are interested in determining whether the parameter θ lies in a set $\Theta_0 \subset \Theta$. One way of phrasing this problem formally, is to say that we would like to test the *null hypothesis*

$$H_0: \quad \theta \in \Theta_0$$

against the alternative

$$H_1: \quad \theta \in \Theta_1 = \Theta \backslash \Theta_0.$$

Much of the formalization of statistical tests is due to Neyman and Pearson, which is why the tests described below are often called Neyman-Pearson tests. In the Neyman-Pearson framework the testing problem is formalized through a decision space $\mathcal{D}$ restricted to $\{1, 0\}$ or {yes, no}, or {accept, reject} or {don't reject, reject}. If Θ_0 (Θ_1) is a singleton $\{\theta_0\}$ ($\{\theta_1\}$) then the null (alternative) hypothesis is a *point* hypothesis, otherwise it is a *composite* hypothesis.

If we let $\mathcal{D} = \{1, 0\}$ then we can view a test as an estimator of the indicator function $\mathbb{I}\{\theta \in \Theta_0\}$ and apply some of the ideas developed in the discussion of point estimation. We

shall denote test procedures by $\varphi(\cdot)$. Test procedures can be evaluated under the following loss function

$$L(\theta,\varphi) = \begin{cases} 1 & \text{if } \varphi \neq \mathbb{I}\{\theta \in \Theta_0\} \\ 0 & \text{otherwise} \end{cases},$$

where $\{\theta \in \Theta_0\}$ is the indicator function that is one if the condition is satisfied and zero otherwise.

Hypothesis testing is, for the most part, a frequentist endeavor. Thus, much of our discussion will focus on the frequentist perspective.

3.1 The Frequentist View

Frequentist Risk, Type-I and II Errors, Power. We can calculate the frequentist risk associated with testing procedure as we did in the case of point estimators

$$R(\theta,\varphi(\cdot)) = \mathbb{E}_\theta^X[L(\theta,\varphi(X))].$$

As before, the frequentist risk is a function of θ. In the context of testing the following terminology is frequently used. The type-I error of a test is the probability of rejecting a null hypothesis if it is correct

$$\text{type-I error} = \alpha(\theta) = R(\theta,\varphi(\cdot)) \quad \theta \in \Theta_0$$

Since the type-I error depends on θ we define the size as the supremum of type-I errors:

$$\text{size} = \sup_{\theta\in\Theta_0} \alpha(\theta).$$

The power of a test is the probability of rejecting a null hypothesis if it is false:

$$\text{power} = \beta(\theta) = 1 - R(\theta,\varphi(\cdot)) \quad \theta \in \Theta_1.$$

The function $1-\beta(\theta)$ is called the type-II error.

The notion of optimality that is commonly used in the testing context differs from our concepts of minimax and admissibility. In general we would like tests that satisfy a constraint on the type-I error and have large power. Before we examine optimality more closely, here are two methods of generating tests.

***t*-Test.** Suppose that $X \sim N(\theta,1)$ and we want to test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0.$$

Here the null hypothesis is a point hypothesis (only one value of θ) and the alternative hypothesis is a composite hypothesis (many values of θ). The idea of a test is to reject if one observes something in the data that is deemed unlikely under the null hypothesis. If $\theta = \theta_0$ then large deviations of X from θ_0 are unlikely. More formally, if H_0 is true, then

$$\hat{\theta} = X \sim N(\theta_0, 1).$$

Thus,

$$Z = \frac{\hat{\theta} - \theta_0}{1} \sim N(0,1).$$

We reject the null hypothesis if $|Z|$ is large. How large? We will calibrate the cut-off value to control the type-I error:

$$\varphi(X) = \mathbb{I}\{|Z(X)| \leq z_{\alpha/2}\},$$

where $z_{\alpha/2}$ satisfies $\Phi_N(z_{\alpha/2}) = 1 - \alpha/2$ and $\Phi_N(\cdot)$ is the cdf of a $N(0,1)$.

Likelihood-Ratio Test. Suppose that $X|\theta \sim N(\theta, 1)$ and we want to test

$$H_0 : \ \theta = \theta_0 \quad \text{vs.} \quad H_1 : \ \theta \neq \theta_0.$$

The likelihood-ratio test is defined as

$$\varphi(x) = \begin{cases} 1 & \text{if } \frac{\sup_{\theta \in \Theta_1} \ p(x|\theta)}{\sup_{\theta \in \Theta_0} \ p(x|\theta)} < k \\ 0 & \text{otherwise} \end{cases}$$

For a point hypothesis under a likelihood function that is continuous in θ, we can simplify the statistic as follows:

$$\varphi(x) = \begin{cases} 1 & \text{if } \frac{p(x|\hat{\theta}_{mle})}{p(x|\theta_0)} < k \\ 0 & \text{otherwise} \end{cases}$$

In our model $\hat{\theta}_{mle} = x$ and the likelihood takes the form

$$p(x|\theta) = (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x-\theta)^2\right\}.$$

Thus,

$$\frac{p(x|\hat{\theta}_{mle})}{p(x|\theta_0)} = \exp\left\{\frac{1}{2}(x-\theta_0)^2\right\}$$

This expression is a bit awkward because of the exponential term. We will take the logarithm and multiply it by 2:

$$LR(X) = 2\ln\left(\frac{p(x|\hat{\theta}_{mle})}{p(X|\theta_0)}\right) = (X-\theta_0)^2.$$

Under the null hypothesis, $X - \theta_0 \sim N(0,1)$, which means that $LR(X) \sim \chi^2(1)$. The test rejects the null hypothesis whenever $LR \geq \kappa$ or $|X| \geq \sqrt{\kappa}$. Using the properties of a normal

distribution the value of κ that guarantees $\alpha = 0.05$ is $\sqrt{\kappa} = 1.96$. The power function can be calculated from the cumulative density function of a non-central χ^2 distribution.

A Basic Optimality Result. We start with a definition of a uniformly most powerful test:

Definition 4 *If $\alpha \in (0,1)$ and $\mathcal{C}_\alpha$ is the class of procedures φ satisfying the following constraint on the type I error*

$$\sup_{\theta_0 \in \Theta} R(\theta, \varphi(\cdot)) \leq \alpha$$

a test procedure φ is said to be uniformly most powerful (UMP) at level α if it maximizes the power $1 - R(\theta, \varphi(\cdot))$ uniformly on Θ_1 in $\mathcal{C}_\alpha$.

In the simplest case, in which null and alternative hypotheses are point hypotheses, the Neyman-Pearson lemma establishes that there exist UMP test procedures and that they are of the form of a *likelihood ratio* test (see above):

$$\varphi(x) = \begin{cases} 1 & \text{if } \frac{p(x|\theta_1)}{p(x|\theta_0)} < k \\ 0 & \text{otherwise} \end{cases}$$

Here k is determined to satisfy the restriction on the type-I error:

$$R(\theta_0, \varphi(\cdot)) = \int (1 - \varphi(x)) p(x|\theta_0) dx = \alpha.$$

The idea of the proof is the following. Suppose that φ^* is an alternative test with

$$\int (1 - \varphi^*(x)) p(x|\theta_0) dx \leq \alpha$$

Denote by S^+ the set in the sample space where $\varphi^*(x) > \varphi(x)$, that is φ^* accepts but φ does not. Likewise, denote by S^- the set in the sample space where $\varphi^*(x) < \varphi(x)$, that is φ accepts but φ^* does not. By construction we have

$$\begin{aligned} p(x|\theta_1) &\geq k p(x|\theta_0) \quad \text{if} \quad x \in S^+ \\ p(x|\theta_1) &< k p(x|\theta_0) \quad \text{if} \quad x \in S^- \end{aligned}$$

Thus,

$$\begin{aligned} \int (\varphi^* - \varphi)[p(x|\theta_1) - k p(x|\theta_0)] dx &= \int_{S^+} (\varphi^* - \varphi)[p(x|\theta_1) - k p(x|\theta_0)] dx \\ &\quad + \int_{S^-} (\varphi - \varphi^*)[k p(x|\theta_0) - p(x|\theta_1)] dx \\ &\geq 0. \end{aligned}$$

The difference in power therefore satisfies

$$\begin{aligned}\int[(1-\varphi)-(1-\varphi^*)]p(x|\theta_1)dx &= \int(\varphi^*-\varphi)p(x|\theta_1)dx \\ &\geq k\int(\varphi^*-\varphi)p(x|\theta_0)dx \\ &\geq k\int[(1-\varphi)-(1-\varphi^*)]p(x|\theta_0)dx \\ &\geq 0.\end{aligned}$$

The last inequality follows from the fact that the type-I error of φ^* is bounded above by α.

Example: Consider our simple location problem: $X \sim N(\theta,1)$ and the null hypothesis $H_0: \theta = 0$ versus $H_1: \theta = \theta_1$. The likelihood ratio takes the form

$$\frac{p(X|\theta_1)}{p(X|0)} = \exp\left\{\frac{1}{2}[X^2-(X-\theta_1)^2]\right\}.$$

It is convenient to use a logarithmic transformation and multiply the log ratio by 2:

$$LR = X^2 - X^2 + 2X\theta_1 - \theta_1^2 = 2X\theta_1 - \theta_1^2.$$

Unfortunately, this test statistic is still a bit awkward. Using another linear transformation, define

$$LR^*(X) = \frac{LR+\theta_1^2}{\theta_1} = X.$$

Thus,

$$\varphi(X) = \begin{cases} 1 & \text{if } X < \kappa \\ 0 & \text{otherwise} \end{cases}$$

Because under the null hypothesis $X \sim N(0,1)$ we can set $\kappa = z_\alpha$ where ζ_α is the one-sided critical value with the property $\Phi_N(z_\alpha) = 1-\alpha$. □

There exist many extensions of the Neyman-Pearson lemma that establish the existence of UMP tests of other hypotheses, such as one-sided hypotheses of the form $H_0: \theta \leq \theta_0$ versus $H_1: \theta > \theta_0$. There also exist UMP tests for certain types of two-sided hypotheses: $H_0: \theta \leq \theta_1$ or $\theta \geq \theta_2$. However, further assumptions on the distribution $p(x|\theta)$ are required. Lehmann's classic text discusses this problem in detail.

p Values. p values could be interpreted as estimators of $\mathbb{I}\{\theta \in \Theta_0\}$. They essentially extend the decision space in a hypothesis testing framework from $\mathcal{D} = \{0,1\}$ to the interval $\mathcal{D} = [0,1]$. Formally, it is the largest type-I error at which H_0 cannot be rejected. Consider the location-shift experiment with $X|\theta \sim N(\theta,1)$. Let the null hypothesis be $H_0: \theta = \theta_0$ and the alternative $H_1: \theta \neq \theta_0$. Recall that our test statistic took the form $Z = (\hat{\theta}-\theta_0)$

which has a $N(0,1)$ distribution under the null hypothesis. Let z^o denote the "observed" value of the statistic, computed from the data. The p value is defined as the tail probability

$$p = 2(1 - \Phi_N(|z^o|)).$$

Note that if $|z^o|$ is large, meaning that the observed value of the statistic lies far in the tails of the sampling distribution $1 - \Phi_N(|z^o|) \approx 0$, i.e., this is evidence against H_0. If $|z^o|$ is small, then $1 - \Phi_N(|z^o|) \approx 0.5$ and the p value is close to 1, i.e., there is evidence in favor of H_0.

3.2 The Bayesian View

Consider the hypothesis test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ where $\Theta_1 = \Theta/\Theta_0$. Hypothesis testing can be interpreted as estimating the value of the indicator function $\{\theta \in \Theta_0\}$. The decision space is 0 ("reject") and 1 ("accept"). Consider the slightly more general loss function

$$L(\theta, \varphi) = \begin{cases} 0 & \varphi = \mathbb{I}\{\theta \in \Theta_0\} \quad \text{correct decision} \\ a_0 & \varphi = 0,\ \theta \in \Theta_0 \quad \text{Type 1 error} \\ a_1 & \varphi = 1,\ \theta \in \Theta_1 \quad \text{Type 2 error} \end{cases} \tag{7}$$

Note that the parameters a_1 and a_2 are part of the econometricians preferences. The optimal decision rule is

$$\delta(Y) = \begin{cases} 1 & P_X^\theta\{\theta \in \Theta_0\} \geq a_1/(a_0 + a_1) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

This can be easily verified. The posterior expected loss is

$$\mathbb{E}_X^\theta[L(\theta, \varphi)] = \mathbb{I}\{\varphi = 0\}a_0\mathbb{P}_X^\theta\{\theta \in \Theta_0\} + \mathbb{I}\{\varphi = 1\}a_1[1 - \mathbb{P}_X^\theta\{\theta \in \Theta_0\}] \tag{9}$$

Thus, one should accept the hypothesis $\theta \in \Theta_0$ (choose $\varphi = 1$) if

$$a_1 P_X^\theta\{\theta \in \Theta_1\} = a_1[1 - \mathbb{P}_X^\theta\{\theta \in \Theta_0\}] \leq a_0\mathbb{P}_X^\theta\{\theta \in \Theta_0\} \tag{10}$$

Often, hypotheses are evaluated according to Bayes factors, that is, the ratio of posterior probabilities and prior probabilities in favor of that hypothesis:

$$B(X) = \frac{\text{Posterior Odds}}{\text{Prior Odds}} = \frac{\mathbb{P}_X^\theta\{\theta \in \Theta_0\}/\mathbb{P}_X^\theta\{\theta \in \Theta_1\}}{\mathbb{P}^\theta\{\theta \in \Theta_0\}/\mathbb{P}^\theta\{\theta \in \Theta_1\}} \tag{11}$$

Note that for testing problems to be interesting, the researcher needs to assign a non-zero prior probability of Θ_0. This means that testing the hypothesis $\theta = 0$ under a continuous prior is not an interesting problem because the resulting continuous posterior will always assign posterior probability zero to the null hypothesis.

Example 1: The parameter space is $\Theta = \{0, 1\}$, and the sample space is $\mathcal{X} = \{0, 1, 2, 3, 4\}$.

	0	1	2	3	4
$\mathbb{P}^X_{\theta=0}$	.75	.140	.04	.037	.033
$\mathbb{P}^X_{\theta=1}$	.70	.251	.04	.005	.004

Suppose that the observed value of X is 2. Note that

$$\begin{aligned}\mathbb{P}^X_{\theta=0}\{X \geq 2\} &= 0.110\\ \mathbb{P}^X_{\theta=1}\{X \geq 2\} &= 0.049\end{aligned}$$

The frequentist interpretation of this result would be that there is significant evidence against $H_0 : \theta = 1$ at the 5 percent level. However, there is not significant evidence against $H_0 : \theta = 0$ at the 10 percent level.

In order to find the Bayesian answer, we need to compute the posterior odds of $\theta = 0$ versus $\theta = 1$. Suppose we consider $\theta = 0$ and $\theta = 1$ as equally likely *a priori*. The posterior probabilities are given by

$$\mathbb{P}^\theta_X\{\theta = j\} = \frac{\mathbb{P}^X_{\theta=j}\{X = 2\}\mathbb{P}^\theta\{\theta = j\}}{\mathbb{P}^X\{X = 2\}}, \quad j = 0, 1.$$

Thus, the posterior odds are

$$\frac{\mathbb{P}^\theta_{X=2}\{\theta = 0\}}{\mathbb{P}^\theta_{X=2}\{\theta = 1\}} = \frac{\mathbb{P}^X_{\theta=0}\{X = 2\}}{\mathbb{P}^X_{\theta=1}\{X = 2\}} = \frac{0.04}{0.04} = 1, \tag{12}$$

implying that the observation $X = 2$ does not favor one versus the other model. □

Example 2: Recall the location-shift model:

$$X|\theta \sim N(\theta, 1), \quad \theta \sim N(0, 1/\lambda)$$

We previously derived the posterior distribution which is given by

$$\theta|X \sim N(\bar{\theta}, V_\theta), \quad \bar{\theta} = \frac{1}{1+\lambda}X, \quad \bar{V}_\theta = \frac{1}{1+\lambda}.$$

Consider the hypothesis $H_0 : \theta < 0$ versus $H_1 : \theta \geq 0$. Then,

$$\mathbb{P}^\theta_X\{\theta < 0\} = \mathbb{P}^\theta_Y\left\{\frac{\theta - \bar{\theta}}{\sqrt{\bar{V}_\theta}} < -\frac{\bar{\theta}}{\sqrt{\bar{V}_\theta}}\right\} = \Phi_N\Big(-\bar{\theta}/\sqrt{\bar{V}_\theta}\Big) \tag{13}$$

where $\Phi_N(\cdot)$ denotes the cdf of a $N(0, 1)$. Suppose that $a_0 = a_1 = 1$ then H_0 is accepted if

$$\Phi\Big(-\bar{\theta}/\sqrt{\bar{V}_\theta}\Big) \geq 1/2 \quad \text{or} \quad \frac{1}{1+\lambda}X < 0. \tag{14}$$

The Classical rule for a one-sided test with a 5 percent significance level is: "accept" H_0 if $X < 1.64$. □

Suppose in the context of Example 2 we would like to test $H_0 : \theta = \theta_0$ versus $H_0 : \theta \neq \theta_0$. Since $\mathbb{P}\{\theta = \theta_0\} = 0$ it follows that $\mathbb{P}_X\{\theta = \theta_0\} = 0$ and the null hypothesis is never accepted. This observations raises the question: are point hypotheses realistic? Only, if one is willing to place positive probability γ on the event that the null hypothesis is true. Consider the modified prior

$$p^*(\theta) = \gamma\delta_{\theta_0}(\theta) + (1-\gamma)p(\theta)$$

where $\delta_0(\theta)$ is a point mass or dirac function. You can think of the function $\delta_x(\theta)$ as the "limit"

$$\lim_{n\longrightarrow\infty} 2n\mathbb{I}\{x - 1/n \leq \theta \leq x + 1/n\}$$

The area under this rectangle is always one. The dirac function has the properties:

(i) $\delta_x(\theta) = \infty$ for $\theta = x$ and zero otherwise.
(ii) It always integrates to one: $\int \delta_x(\theta)d\theta = 1$.
(iii) For $|f(\theta)| \leq M \leq \infty$ it follows that $f(\theta)\delta_x(\theta) = f(x)\delta_x(\theta)$.
(iv) As a consequence $\int f(\theta)\delta_x(\theta)d\theta = f(x)$.

We use the $*$ subscript to denote priors, posteriors, and marginal data densities that are obtained as mixtures of a discrete and continuous distribution. Densities without a $*$ subscript are purely continuous.

Bayes Theorem implies that

$$p_*(\theta|x) = \frac{p(x|\theta)p_*(\theta)}{\int p(x|\theta)p_*(\theta)d\theta}$$

The denominator takes the following form

$$\begin{aligned}\int p(x|\theta)p_*(\theta)d\theta &= \gamma\int p(x|\theta_0)\delta_{\theta_0}(\theta)d\theta + (1-\gamma)\int p(x|\theta)p(\theta)d\theta \\ &= \gamma p(x|\theta_0) + (1-\gamma)\int p(x|\theta)p(\theta)d\theta \\ &= \gamma p(x|\theta_0) + (1-\gamma)p(x).\end{aligned}$$

Thus, we can write the posterior density as

$$\begin{aligned}p_*(\theta|x) &= \frac{\gamma p(x|\theta_0)\delta_{\theta_0}(\theta) + (1-\gamma)p(x|\theta)p(\theta)}{\gamma p(x|\theta_0) + (1-\gamma)p(x)} \\ &= \frac{\gamma p(x|\theta_0)}{\gamma p(x|\theta_0) + (1-\gamma)p(x)}\delta_{\theta_0}(\theta) + \frac{(1-\gamma)p(x)}{\gamma p(x|\theta_0) + (1-\gamma)p(x)}\frac{p(x|\theta)p(\theta)}{p(x)} \\ &= \frac{\gamma p(x|\theta_0)}{\gamma p(x|\theta_0) + (1-\gamma)p(x)}\delta_{\theta_0}(\theta) + \frac{(1-\gamma)p(x)}{\gamma p(x|\theta_0) + (1-\gamma)p(x)}p(\theta|x).\end{aligned}$$

The posterior probability of $\theta = 0$ is given by

Table 1: Posterior Odds in Gaussian Location Model ($\gamma = 1/2$, $\lambda = 1$)

X	0.0	0.5	1.0	1.1	1.64	1.95
Odds $\theta = 0$ vs. $\theta \neq 0$	1.41	1.32	1.10	0.99	0.72	0.55

$$
\begin{aligned}
&\mathbb{P}^{\theta}_{*X}\{\theta = 0\} \\
&= \lim_{\epsilon \longrightarrow 0} \mathbb{P}^{\theta}_{*X}\{0 \leq \theta \leq \epsilon\} \qquad (15)\\
&= \lim_{\epsilon \longrightarrow 0} \frac{\gamma p(x|\theta_0)}{\gamma p(x|\theta_0) + (1-\gamma)p(x)} \int_{\theta_0}^{\theta_0+\epsilon} \delta_{\theta_0}(\theta)d\theta + \frac{(1-\gamma)p(x)}{\gamma p(x|\theta_0) + (1-\gamma)p(x)} \int_{\theta_0}^{\theta_0+\epsilon} p(\theta|x)d\theta \\
&= \frac{\gamma p(x|\theta_0)}{\gamma p(x|\theta_0) + (1-\gamma)p(x)}.
\end{aligned}
$$

Thus, we can rewrite the posterior density as

$$p_*(\theta|x) = \mathbb{P}^{\theta}_{*X}\{\theta = \theta_0\}\delta_{\theta_0}(\theta) + (1 - \mathbb{P}^{\theta}_{*X}\{\theta = \theta_0\})p(\theta|x). \qquad (16)$$

It is a mixture of a pointmass at θ_0 and the continuous posterior distribution $p(\theta|x)$.

The posterior odds are

$$\frac{P^{\theta}_{*X}\{\theta = \theta_0\}}{P^{\theta}_{*X}\{\theta \neq 0\}} = \frac{\gamma}{1-\gamma} \cdot \frac{p(x|\theta_0)}{\int p(x|\theta)p(\theta)d\theta}. \qquad (17)$$

In our Gaussian location model, note that we can write

$$X = \theta + U, \quad \theta \sim N(0, 1/\lambda), \quad U \sim N(0,1).$$

Because the sum of two normal random variables is normally distributed and θ and U are independent, we obtain

$$X \sim N(\theta, 1 + 1/\lambda).$$

Thus for $\theta_0 = 0$,

$$
\begin{aligned}
\frac{P^{\theta}_{*X}\{\theta = 0\}}{P^{\theta}_{*X}\{\theta \neq 0\}} &= \frac{\gamma}{1-\gamma} \cdot \frac{(2\pi)^{-1/2}\exp\left\{-\frac{1}{2}X^2\right\}}{(2\pi)^{-1/2}\left(1+\frac{1}{\lambda}\right)^{-1/2}\exp\left\{-\frac{1}{2(1+1/\lambda)}X^2\right\}} \\
&= \frac{\gamma}{1-\gamma}\left(1+\frac{1}{\lambda}\right)^{1/2}\exp\left\{-\frac{1}{2}\left(1-\frac{\lambda}{1+\lambda}\right)X^2\right\}.
\end{aligned}
$$

Table 1 reports posterior odds for various values of X, assuming that the prior odds are one and $\lambda = 1$. Note that for $X = 1.1$ the odds favor the alternative hypothesis. A frequentist could reject at the 10% level for $X = 1.64$ and at the 5% level for $X = 1.96$.

Note that if we hold X fixed and let $\lambda \longrightarrow 0$ then the posterior odds in favor of $\theta = 0$ diverge to infinity. The intuition is that under the alternative model, the variance of the marginal density goes to infinity, which means that we should expect to observe very extreme realizations of X. To the extent that we don't, we interpret this as evidence for H_0.

4 Coverage Sets

In practice more useful than tests (and point estimators) are coverage sets.

4.1 Frequentist Confidence Sets

In a frequentist world confidence sets $C(X) \subseteq \Theta$ are random sets that are supposed to cover the parameter θ with a pre-specified probability:

$$\inf_{\theta \in \Theta} P_\theta^X[\{\theta \in C(X)\}] \geq 1 - \alpha$$

From a frequentist perspective, X and $C(X)$ are random and θ is fixed. Note that the coverage probability has to be guaranteed for all values of $\theta \in \Theta$ because we do not know θ.

Confidence sets can be constructed from the acceptance region of a frequentist hypothesis test. Suppose we index our hypothesis test $\varphi(\cdot)$ not just by the observation x but also by the value θ_0 associated with the (point) null hypothesis: $\varphi(x, \theta_0)$. Then, consider the set of θ_0's such that the null hypothesis $\theta = \theta_0$ is accepted given X:

$$C(X) = \{\theta_0 \in \Theta \,|\, \varphi(X, \theta_0) = 1\}$$

such that

$$\mathbb{I}\{\theta_0 \in C(X)\} = \varphi(X, \theta_0).$$

In turn (dropping the 0 subscript)

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta^X[\{\theta \in C(X)\}] = \inf_{\theta \in \Theta} \mathbb{E}_\theta^X[\varphi(X, \theta)] = 1 - \sup_{\theta \in \Theta}(1 - \mathbb{E}_\theta^X[\varphi(X, \theta)]) \geq 1 - \alpha$$

because we started from a test that has a type-I error bounded by α. In general, we want the confidence sets to have a small volume subject to a constraint on the coverage probability. In a nutshell: powerful tests lead to small confidence sets.

Example: In our simple location model $X \sim N(\theta, 1)$ we can invert the likelihood ratio test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, which leads to the familiar confidence set $C(x) = [x - 1.96,\ x + 1.96]$. $\square$

The notion of inverting a test to construct a confidence interval becomes more powerful in non-standard testing problems in which the confidence interval does not simply take the form of "point estimate plus/minus two-times the standard error estimate."

Example: Consider the problem $X|\theta \sim N(\theta, 1)$ with the additional constraint $\theta \geq 0$. A confidence set for θ can be obtained by inverting the test of $H_0 : \theta = \tilde{\theta}$. The likelihood ratio statistic for this test is given by

$$LR = 2\left[-\frac{1}{2}(X - \max\{0, X\})^2 + \frac{1}{2}(X - \tilde{\theta})^2\right] = (X - \tilde{\theta})^2 - (X - \max\{0, X\})^2.$$

To obtain the critical value for the LR statistic, we need to derive its distribution under the null hypothesis. Let's write $X = \tilde{\theta} + U$, where $U \sim N(0, 1)$. We have to distinguish the case $X \geq 0$ versus $X < 0$, which translates into $U \geq -\tilde{\theta}$ versus $U < -\tilde{\theta}$. Now write

$$LR \sim \begin{cases} U^2 & \text{if } U \geq -\tilde{\theta} \\ -\tilde{\theta}^2 - 2\tilde{\theta}U & \text{otherwise} \end{cases} \tag{18}$$

For small values of $\tilde{\theta}$ the distribution of the test statistic and its critical value differ from the standard χ^2 distribution. You can construct the confidence interval as follows.

Choose a grid $\mathcal{T}$ of θ values. For each $\tilde{\theta} \in \mathcal{T}$

1. Compute the observed value $LR^o(\tilde{\theta})$ of the likelihood ratio test statistic.
2. Simulate the distribution of LR (by generating *iid* random draws from $U \sim N(0, 1)$ and evaluating the right-hand-side of (18) and determine the critical value for a size α test as $1 - \alpha$ quantile of the simulated distribution of LR. Call it $CV(\tilde{\theta})$
3. If $LR^o(\tilde{\theta}) \leq CV(\tilde{\theta})$ then include $\tilde{\theta}$ into the confidence set. □

4.2 Bayesian Credible Sets

For comparison, a Bayesian credible set is defined as

$$\mathbb{P}_X^{\theta}[\{\theta \in C(X)\}] \geq 1 - \alpha.$$

Here X is fixed and θ is random. Be mindful of the difference in interpretation! There are many types of credible sets that one can construct from a posterior distribution. The smallest (in terms of volume) are the so-called highest-posterior-density (HPD) sets:

$$C_X = \{\theta : p(\theta|X) \geq k_\alpha\}$$

where k_α is the largest bound such that

$$P_X^{\theta}\{\theta \in C_X\} \geq 1 - \alpha.$$

References

My exposition is based on Robert (2007). Lehmann (and Casella, 1998) is the classic reference for the theory of point estimation. Lehmann (and Romano, 2005) is the classic reference for the frequentist testing theory. The book is much older than the publication date of the latest version suggests.

Casella, George and Roger Berger (2001): *Statistical Inference*, Duxbury Press.

Lehmann, Erich and George Casella (1998): *Theory of Point Estimation*, Springer Verlag.

Lehmann, Erich and Joseph Romano (2005): *Testing Statistical Hypotheses*, Springer Verlag.

Robert, Christian (2007): *The Bayesian Choice*, Springer Verlag.